

Moral Lineage Tracing

Florian Jug^{1,*}, Evgeny Levinkov^{2,*}, Corinna Blasse¹, Eugene W. Myers¹, Bjoern Andres^{2,†}

Abstract

Lineage tracing, the tracking of living cells as they move and divide, is a central problem in biological image analysis. Solutions, called lineage forests, are key to understanding how the structure of multicellular organisms emerges. We propose an integer linear program (ILP) whose feasible solutions define a decomposition of each image in a sequence into cells (segmentation), and a lineage forest of cells across images (tracing). Unlike previous formulations, we do not constrain the set of decompositions, except by contracting pixels to superpixels. The main challenge, as we show, is to enforce the morality of lineages, i.e., the constraint that cells do not merge. To enforce morality, we introduce path-cut inequalities. To find feasible solutions of the NP-hard ILP, with certified bounds to the global optimum, we define efficient separation procedures and apply these as part of a branch-and-cut algorithm. We show the effectiveness of this approach by analyzing feasible solutions for real microscopy data in terms of bounds and run-time, and by their weighted edit distance to ground truth lineage forests traced by humans.

1 Introduction

Phenomenal progress in microscopy allows biologists to image large numbers of living cells as they move and divide [28, 41]. Such observations are essential in developmental biology for studying embryogenesis and tissue formation [29, 31, 34]. Consequently, the tracing of cells and their lineages in sequences of images has become a central problem in biological image analysis [4, 5, 27].

The lineage tracing problem comprises two sub-problems, cf. Fig. 1: The first sub-problem is to identify the cells in every individual image. The second sub-problem is to connect every cell identified in an image to the same cell and descendant cells identified in successive images. Solutions are constrained by the fact that every cell has precisely one direct progenitor (predecessor), i.e., cells do not merge. Exceptions are image boundaries, where cells can enter the field without their direct progenitors being visible. Moreover, cells never split into more than two cells, unless the temporal resolution is too low to separate divisions. Last but not least, cells can disappear by leaving the field of view.

The first sub-problem is an image decomposition problem: If every pixel shows part of a cell and no pixel shows background, the objective is to decompose the pixel grid graph into precisely one component per cell. If pixels potentially show background, the objective is to jointly select and decompose a subgraph of the pixel grid graph such that there is precisely one component for each cell and no component for the background.

The second sub-problem is a reconstruction problem: Here, the objective is to identify a set of pairwise disjoint lineage trees (depicted in Fig. 1 in red and green) whose nodes are cells.

It is understood that errors in the image decomposition make it harder to reconstruct the true lineage forest. Yet, attempts at reconstructing the lineage forest can

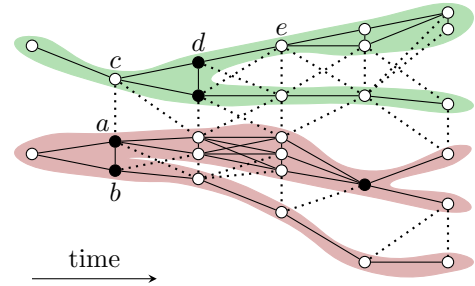


Figure 1 Given a sequence of images, taken at successive points in time, and a decomposition of each image into *cell fragments* (depicted above as nodes), the objective of lineage tracing is to join fragments of the same cell within and across images, e.g. $\{a, b\}$ and $\{c, d\}$, and fragments of descendant cells across images, e.g. $\{d, e\}$. Joins (cuts) are depicted as solid (dashed) lines. Fragments of dividing cells are depicted as black nodes.

help to avoid such errors. Therefore, we state a joint optimization problem whose feasible solutions define, for every image, a decomposition (segmentation) into cells and, across images, a lineage forest. Unlike in prior work, we do not constrain the set of decompositions, except by contracting pixels to superpixels.

2 Related Work

The image decomposition problem has been tackled in the form of various optimization problems, including the minimum cost spanning forest problem, i.e., agglomerative clustering [1], balanced cut problems, i.e., spectral clustering [10, 39], and the minimum cost multicut problem, i.e., correlation clustering [32]. We build on its formulation as a minimum cost multicut problem, an optimization problem studied in [16, 18] which is NP-hard [12, 17] and has been used for image segmentation in [3, 7, 8, 9, 11, 13, 14, 23, 24, 25, 30, 32, 48, 47].

The lineage forest reconstruction problem has been cast as an optimization problem in [20, 22, 26, 35, 37, 42, 43, 44, 45]. If cells neither die nor appear or disap-

¹MPI of Molecular Cell Biology and Genetics, Dresden

²MPI for Informatics, Saarbrücken

*Contributed equally

[†]Correspondence: andres@mpi-inf.mpg.de

pear at the boundary of the image and if one drops the constraint that cells split into at most two direct descendant cells, e.g. because temporal image resolution is too low to separate divisions, the problem can be formulated as a minimum cost k disjoint arborescence problem [38, Section 53.9], as shown in [44, 45]. Here, k is the number of cells visible in the first image. This problem can be solved in strongly polynomial time [19]. With the additional constraint that cells split into at most two descendant cells, the problem becomes NP-hard, and so do generalizations [20, 22, 26, 35, 37, 42, 43] that model, e.g., (dis)appearance.

One lineage tracing approach [26] copes with imperfect decompositions by over-segmenting individual images. This guarantees that every cell is represented by at least one component and that every component represents at most one cell. One advantage is feasibility, i.e., the fact that the true lineage forest is represented by at least one set of disjoint arborescences. One disadvantage is the loss of robustness: While, for the true decomposition, every component belongs to precisely one arborescence and thus, every error in the set of disjoint arborescences implies at least one second error, which renders solutions robust to perturbations of the objective function, for over-segmentations, this property is lost. Another disadvantage is the fact that the number of progenitor cells is not determined by the number of components of the first image. Over-estimates result in excessive arborescences that typically conflict with correct ones. Under-estimates result in a loss of lineage trees. As in [26], we consider an over-decomposition of each image into cell fragments. In contrast to [26] where each node of a lineage forest is a single representative fragment, each node in the lineage forests we consider is a clusters of fragments. This idea of clustering instead of selecting is used in [40] to track multiple people in a video. The optimization problem defined there is a hybrid of a minimum cost multicut problem and a disjoint path problem. The optimization problem we propose is a hybrid of a minimum cost multicut problem and a disjoint arborescence problem.

Two techniques have been proposed to deal with over and under-decomposition simultaneously: The first [37] is to allow single image components to represent multiple cells and thus be part of multiple lineages. This relaxation of the disjointness constraint of the arborescence problem introduces additional feasible solutions that can represent the true lineage forest even in the presence of under-decomposition. The same idea is used in [43] for the reconstruction of curvilinear structures and in [46] for the tracking of objects in containers. The second technique [20, 22, 36] considers a hierarchy of alternative decompositions and casts lineage tracing as an optimization problem whose feasible solutions select and connect components from the hierarchy. Constraints guarantee that selected components are mutually consistent and consistent with a set of disjoint lineages. As in [20, 22], the feasible solutions we propose define (i) for every image a decomposition into cells and, (ii) a lineage forest across images. In contrast to prior work, we do not constrain the set of decompositions, except by

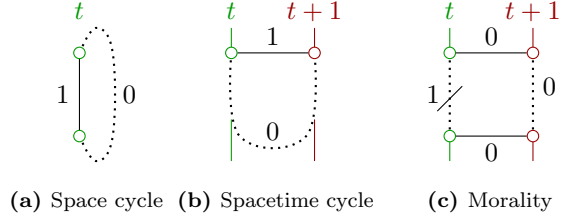


Figure 2 Depicted above are examples (graphs and 01-labelings of edges) in which inequalities (1)-(3) are violated. **(a)** An inequality (1) is violated iff there exist $t \in \mathbb{N}$ and a cycle Y in G_t in which precisely one edge is labeled 1. **(b)** An inequality (2) is violated iff there exist $t \in \mathbb{N}$, an edge $\{v, w\} \in E_{t,t+1}$ labeled 1 and a path in G_t^+ connecting v to w in which all edges are labeled 0. **(c)** An inequality (3) is violated iff there exist $t \in T$ and nodes $v_t, w_t \in V_t$ and $v_{t+1}, w_{t+1} \in V_{t+1}$ connected by edges $\{v_t, v_{t+1}\}, \{w_t, w_{t+1}\} \in E_{t,t+1}$ labeled 0, such that v_t and w_t are separated by a cut in G_t with all edges labeled 1 and v_{t+1} and w_{t+1} are connected by a path in G_{t+1} with all edges labeled 0.

contracting pixels to superpixels. We compare our experimental results to a state-of-the-art software system for lineage tracing [2].

Unlike in the work discussed above, feasible lineages and costs of feasible lineages can be defined recursively, as is done in *particle filtering*. Cf. [15] for a recent comprehensive comparison and [5] for a recent application to lineage tracing.

3 Optimization Problem

In this section, lineage tracing is cast as an optimization problem. In Section 3.1 we introduce the set of feasible solutions of the moral lineage tracing problem. The objective function and optimization problem are defined in Section 3.2.

3.1 Feasible Set

In order to encode a combinatorial number of feasible solutions, we define *hypothesis graphs*. In a hypothesis graph, each node corresponds to one superpixel of one image in a sequence and is referred to as a *cell fragment* (Def. 1). In order to encode a single feasible solution, we define *lineage graphs* (Def. 2). A lineage graph is a subgraph of the hypothesis graph that defines, within each image, a clustering of cell fragments into cells and, across images, a lineage forest of cells (Lemma 1). In order to state an optimization problem whose feasible solutions are lineage graphs in the form of an ILP, we identify the characteristic functions of lineage graphs with 01-labelings of the edges of the hypothesis graph that satisfy a system of linear inequalities (Lemma 2).

Definition 1 A *hypothesis graph* is a node-labeled graph¹ $G = (V, E, \tau)$ in which every edge $\{v, w\} \in E$ holds $|\tau(v) - \tau(w)| \leq 1$. Any $v \in V$ is called a (*cell*) *fragment*, and $\tau(v)$ is called its *time index*.

¹All graphs are assumed to be finite, simple and undirected. A node labeling of a graph (V, E) is a map $\tau : V \rightarrow \mathbb{N}$.

The intuition is this: For any distinct fragments $v, w \in V$ with $\tau(v) = \tau(w)$, the presence of the edge $\{v, w\} \in E$ indicates the possibility that v and w are fragments of the same cell. For any fragments $v, w \in V$ with $\tau(w) = \tau(v) + 1$, the presence of the edge $\{v, w\} \in E$ indicates the possibility that v and w are fragments of the same cell, observed at successive points in time, as well as the possibility that v is a fragment of a progenitor cell of the cell of w .

Next, we characterize those subgraphs of a hypothesis graph that we consider as feasible solutions. For clarity, we propose some notation: For every $t \in \mathbb{N}$, let $V_t := \tau^{-1}(t)$ the set of all fragments having the time index t . Let $G_t = (V_t, E_t)$ the subgraph of G induced by V_t . Let $E_{t,t+1} := \{\{v, w\} \in E \mid v \in V_t \wedge w \in V_{t+1}\}$ the set of those edges of G that connect a fragment v having the time index t to a fragment w having the time index $t + 1$. Let $G_t^+ = (V_t^+, E_t^+)$ the subgraph of G induced by $V_t^+ := V_t \cup V_{t+1}$. Finally, let $V_{\geq t} := \cup_{t'=t}^{\infty} V_{t'}$ the fragments of at least time index t , and $E_{\geq t} := \cup_{t'=t}^{\infty} (E_{t'} \cup E_{t',t'+1})$ the set of all edges of G between such fragments.

Definition 2 For any hypothesis graph $G = (V, E, \tau)$, a set $C \subseteq E$ is called a *lineage cut* of G , and (V, \bar{C}) with $\bar{C} := E \setminus C$ is called a *lineage (sub)graph* of G , iff the following conditions hold:

1. For any $t \in \mathbb{N}$, the set $E_t \cap C$ is a multicut² of G_t
2. For any $t \in \mathbb{N}$ and any $\{v, w\} \in E_{t,t+1} \cap C$, v and w are not connected by any path in the graph $(V_t^+, E_t^+ \cap \bar{C})$
3. For any $t \in \mathbb{N}$, any $v_t, w_t \in V_t$ and $v_{t+1}, w_{t+1} \in V_{t+1}$ such that $\{v_t, v_{t+1}\} \in E \cap \bar{C}$ and $\{w_t, w_{t+1}\} \in E \cap \bar{C}$, and for any path in $(V, E_{t+1} \cap \bar{C})$ from v_{t+1} to w_{t+1} , there exists a path in $(V, E_t \cap \bar{C})$ from v_t to w_t .

If these conditions are satisfied then, for any $t \in \mathbb{N}$ and any non-empty, maximal connected subgraph (V'_t, E'_t) of $(V_t, E_t \cap \bar{C})$, its node set V'_t is called a *cell* at time index t .

An intuition for Conditions 1–3 is offered by Lemma 1 and the proof.

Lemma 1 For every $t \in \mathbb{N}$, a lineage graph well-defines a decomposition of G_t whose components are the cells at time index t . Across time, a lineage graph well-defines a (lineage) forest of cells.

PROOF Condition 1 guarantees that every subgraph defining a cell is node-induced, i.e., for every $t \in \mathbb{N}$ and every $\{v, w\} \in E_t$: $\{v, w\} \in \bar{C}$ iff v and w are fragments of the same cell. Condition 2 guarantees, for every $t \in \mathbb{N}$, every cell V'_t at time index t , and every cell V'_{t+1} at time index $t + 1$ that either all edges of G between V'_t and V'_{t+1} are in \bar{C} , or none. Condition 3 guarantees, for every $t \in \mathbb{N}$ and every distinct cells V'_t, V''_t at time index t that these are not connected in $(V, E_t^+ \cap \bar{C})$ to the same cell at time index $t + 1$. This guarantees, by induction, that V'_t, V''_t are not connected by any path in the graph $(V_{\geq t}, E_{\geq t} \cap \bar{C})$. This guarantees that distinct cells never merge. \square

²A multicut of $G_t = (V_t, E_t)$ is a subset $M \subseteq E_t$ such that, for every cycle Y in G_t : $|M \cap C| \neq 1$ [16].

Lemma 2 For any hypothesis graph $G = (V, E, \tau)$ and any $x \in \{0, 1\}^E$, the set $x^{-1}(1)$ of edges labeled 1 is a lineage cut of G iff Conditions (1)–(3) hold. It is sufficient in (1) to consider only chordless cycles.

$$\forall t \in \mathbb{N} \forall Y \in \text{cycles}(G_t) \forall e \in Y :$$

$$x_e \leq \sum_{e' \in Y \setminus \{e\}} x_{e'} \quad (1)$$

$$\forall t \in \mathbb{N} \forall \{v, w\} \in E_{t,t+1} \forall P \in vw\text{-paths}(G_t^+) :$$

$$x_{vw} \leq \sum_{e \in P} x_e \quad (2)$$

$$\forall t \in \mathbb{N} \forall \{v_t, v_{t+1}\}, \{w_t, w_{t+1}\} \in E_{t,t+1}$$

$$\forall T \in v_t w_t\text{-cuts}(G_t) \forall P \in v_{t+1} w_{t+1}\text{-paths}(G_{t+1}) :$$

$$1 - \sum_{e \in T} (1 - x_e) \leq x_{v_t v_{t+1}} + x_{w_t w_{t+1}} + \sum_{e \in P} x_e \quad (3)$$

The set of all $x \in \{0, 1\}^E$ that satisfy (1)–(3) is denoted by X_G . Examples of violated inequalities are depicted in Fig. 2. Note that the path-cut inequalities (3) guarantee that any fragments joined to the same cell at time $t + 1$ cannot be joined to fragments of distinct cells at time t , i.e., morality. Complementary to the proof given below, a discussion of (dis)connectedness w.r.t. a multicut can be found in [6].

PROOF If Condition 1 in Def. 2 holds for a set $C \subseteq E$, then all inequalities (1) are satisfied by the $x \in \{0, 1\}^E$ such that $x^{-1}(1) = C$. Otherwise, there would exist a $t \in \mathbb{N}$, a cycle Y of G_t and an $e \in Y$ such that $x_e = 1$ and $\forall e' \in Y \setminus \{e\}$: $x_{e'} = 0$. This implies $|Y \cap C| = 1$, in contradiction to the assumption that $C \cap E_t$ is a multicut of G_t . Conversely, if all inequalities (1) are satisfied by an $x \in \{0, 1\}^E$, then $C := x^{-1}(1)$ satisfies Condition 1 in Def. 2. Otherwise, there would exist a $t \in \mathbb{N}$ for which $C \cap E_t$ is not a multicut of G_t . Thus, there would exist a cycle Y of G_t and an $e \in Y$ such that $Y \cap C = \{e\}$, by definition of a multicut. Hence, the inequality (1) for that cycle Y and that edge e of Y would be violated by x . The sufficiency of chordless cycles follows from (1) and is established, e.g., in [16].

If Condition 2 in Def. 2 holds for a set $C \subseteq E$, then all inequalities (2) are satisfied by the $x \in \{0, 1\}^E$ such that $x^{-1}(1) = C$. Otherwise, there would exist $t \in \mathbb{N}$, $\{v, w\} \in E_{t,t+1}$ and a path $P \in vw\text{-paths}(G_t^+)$ such that $x_{vw} = 1$ and $x_P = 0$. From $x_{vw} = 1$ follows $\{v, w\} \in E_{t,t+1} \cap C$. From $x_P = 0$ follows that v and w are connected by P in $(V_t^+, E_t^+ \cap \bar{C})$. Both statements together contradict the assumption. Conversely, if all inequalities (2) are satisfied by an $x \in \{0, 1\}^E$, then $C := x^{-1}(1)$ satisfies Condition 2 in Def. 2. Otherwise, there would exist $t \in \mathbb{N}$, $\{v, w\} \in E_{t,t+1} \cap C$ and a path $P \in vw\text{-paths}(V_t^+, E_t^+ \cap \bar{C})$. From this follows $x_{vw} = 1$ and $x_P = 0$, in contradiction to the assumption that (2) is satisfied.

If Condition 3 in Def. 2 holds for a set $C \subseteq E$, then all inequalities (3) are satisfied by the $x \in \{0, 1\}^E$ such that $x^{-1}(1) = C$. Otherwise, there would exist $t \in \mathbb{N}$, $v_t, w_t \in V_t$, $v_{t+1}, w_{t+1} \in V_{t+1}$, a path $P \in v_{t+1} w_{t+1}\text{-paths}(G_{t+1})$, and a cut $T \in v_t w_t\text{-cuts}(G_t)$

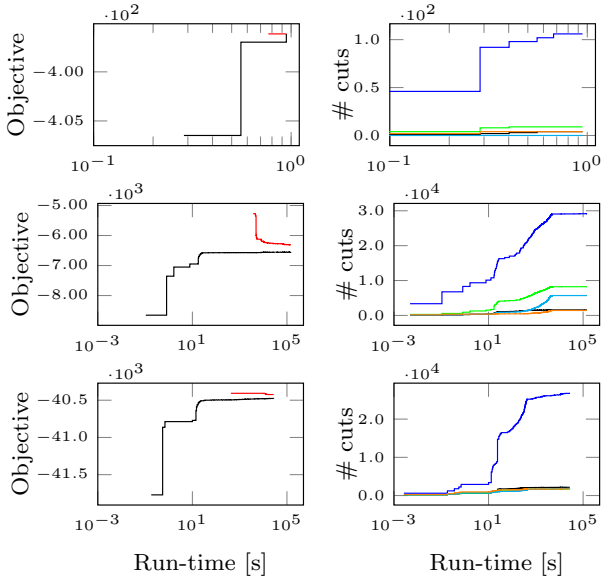


Figure 3 Depicted above is the convergence of the branch-and-cut algorithm for three instance of the moral lineage tracing problem. These are, from top to bottom, HeLa-small, HeLa-test and Flywing. Graphs on the left show the objective values of intermediate integer feasible solutions (—) and lower bounds (---). It can be seen from these graphs that the first problem is solved to optimality. Graphs on the right show numbers of cuts: morality (—), spacetime cycle (—), space cycle (—), appearance (—), and disappearance (—). It can be seen from these graphs that violated morality cuts dominate.

such that $x_{v_t, v_{t+1}} = 0$ and $x_{w_t, w_{t+1}} = 0$ and $x_P = 0$ and $x_T = 1$. P witnesses the existence of a $v_{t+1}w_{t+1}$ -path in $(V, E_{t+1} \cap \bar{C})$. The existence of T certifies the non-existence of a $v_t w_t$ -path in $(V, E_t \cap \bar{C})$. Both statements together contradict the assumption. Conversely, if all inequalities (3) are satisfied by an $x \in \{0, 1\}^E$, then $C := x^{-1}(1)$ satisfies Condition 3 in Def. 2. Otherwise, there would exist $t \in \mathbb{N}$, $v_t, w_t \in V_t$ and $v_{t+1}, w_{t+1} \in V_{t+1}$ such that $\{v, w\} \in E_{t,t+1} \cap \bar{C}$ and $\{v_{t+1}, w_{t+1}\} \in E_{t,t+1} \cap \bar{C}$ and such that there exist $P \in v_{t+1}w_{t+1}$ -paths($V_{t+1}, E_{t+1} \cap \bar{C}$) and $T \in v_t w_t$ -cuts($V_t, E_t \cap \bar{C}$). Hence, $x_{v_t, v_{t+1}} = 0$ and $x_{w_t, w_{t+1}} = 0$ and $x_P = 0$ and $x_T = 1$, in contradiction to the assumption that (3) is satisfied. \square

3.2 Objective Function

Definition 3 A *priced hypothesis graph* is a tuple $(V, E, \tau, c, c^+, c^-)$ with (V, E, τ) a hypothesis graph, $c : E \rightarrow \mathbb{R}$ and $c^+, c^- : V \rightarrow \mathbb{R}_0^+$. For any $e \in E$, c_e is called the *cut cost* of e . For any $v \in V$, c_v^+ and c_v^- are called the *appearance* and *disappearance cost* of v , respectively.

Below, the ILP we propose is defined w.r.t. a priced hypothesis graph $G = (V, E, \tau, c, c^+, c^-)$. This ILP has the following properties: Every feasible solutions defines a lineage subgraph of G . For every $\{v, w\} = e \in E$, the objective function assigns the cost (or reward) c_e to all lineage graphs in which the cell fragments v and w belong to distinct cells. For every $t \in \mathbb{N}$ and

every $v \in V_{t+1}$, the objective function assigns the non-negative (appearance) cost c_v^+ to all lineage graphs in which the fragment v is not joined with any fragment in V_t . For every $t \in \mathbb{N}$ and every $v \in V_t$, the objective function assigns the non-negative (disappearance) cost c_v^- to all lineage graphs in which the fragment v is not joined with any fragment in V_{t+1} .

Definition 4 For any priced hypothesis graph $G = (V, E, \tau, c, c^+, c^-)$, the instance of the *moral lineage tracing problem* w.r.t. G is the ILP in $x \in \{0, 1\}^E$ and $x^+, x^- \in \{0, 1\}^V$ written below.

$$\min_{x, x^+, x^-} \sum_{e \in E} c_e x_e + \sum_{v \in V} c_v^+ x_v^+ + \sum_{v \in V} c_v^- x_v^- \quad (4)$$

$$\text{subject to } x \in X_G \quad (5)$$

$$\forall t \in \mathbb{N} \forall v \in V_{t+1} \forall T \in V_t v\text{-cuts}(G_t^+) :$$

$$1 - x_v^+ \leq \sum_{e \in T} (1 - x_e) \quad (6)$$

$$\forall t \in \mathbb{N} \forall v \in V_t \forall T \in v V_{t+1}\text{-cuts}(G_t^+) :$$

$$1 - x_v^- \leq \sum_{e \in T} (1 - x_e) \quad (7)$$

If, in an inequality of (6), all edges in the cut T are labeled 1, then $x_v^+ = 1$. Otherwise, for every feasible solution x the same solution but with $x_v^+ := 0$ is not worse (as $0 \leq c_v^+$, by definition of c^+). Thus, a cost $c_v^+ \neq 0$ is payed iff fragment v appears at time $t + 1$. The argument for (7) is analogous.

4 Optimization Algorithm

4.1 Efficient Separation Procedures

Below, we define, for each class of inequalities, (1)–(3), (6) and (7), a procedure that takes any (x, x^+, x^-) as the input. If any inequality is violated, it terminates and outputs at least one of these. If no inequality is violated, it terminates and outputs the empty set. These separation procedures are used in a branch-and-cut algorithm described in the next section. They are also used in the preparation of the experiments described in Section 5, to certify the well-definedness of lineages we traced manually. Below, the term “complexity” means worst-case computational time complexity.

To separate infeasible solutions by inequalities (1) for a given t , we label maximal subgraphs of G_t connected by edges labeled 0. Then, for every $\{v, w\} = e \in E_t$ with $x_e = 1$ and with v and w being in the same subgraph, we search for a shortest vw -path P in G_t such that $x_P = 0$, using breadth-first-search (BFS). If the path is chordless, we output the inequality defined by the cycle $P \cup \{e\}$ and e . The complexity is $O(|V_t||E_t|)$.

To separate infeasible solutions by inequalities (2) for a given t , we label maximal subgraphs of G_t^+ connected by edges labeled 0. Then, for every $\{v, w\} = e \in E_{t,t+1}$ with $x_e = 1$ and with v and w being in the same subgraph, we search for a shortest vw -path P in G_t^+ such that $x_P = 0$ using BFS. We output the inequality defined by the cycle $Y := P \cup \{e\}$ and $e \in Y$. The complexity is $O(|V_t^+||E_{t,t+1}|)$.

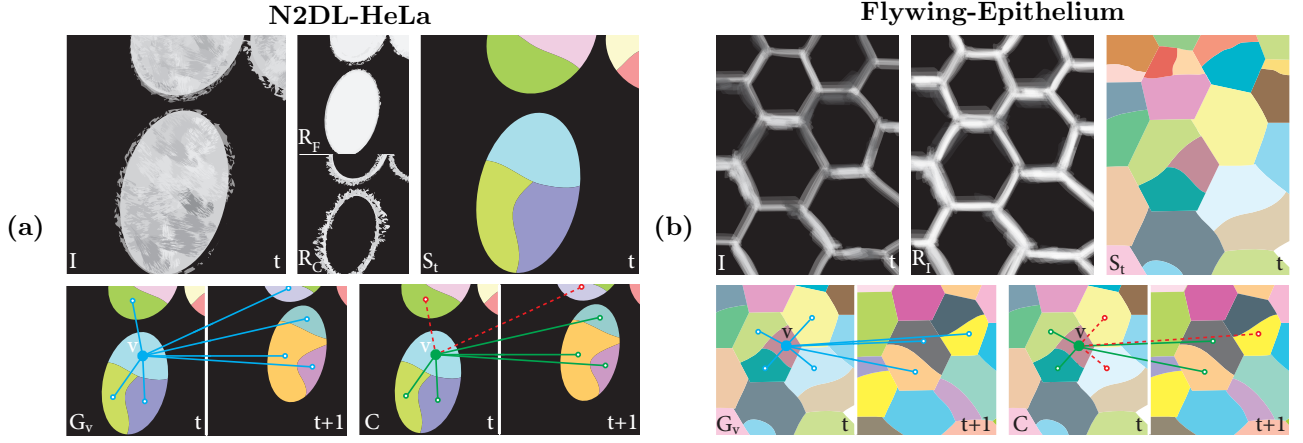


Figure 4 Sketched above is the construction of an instance of the MLT for (a) the N2DL-HeLa Data and (b) the Flying-Epithelium Data. For every time index t and the respective image I in the sequence, foreground probabilities P_F of pixels depicting a cell and probabilities P_C and P_I of pixels depicting a boundary are estimated and used to remove the background and to decompose the image into cell fragments S_t . A hypothesis graph (depicted for one fragment in blue) connects nearby cell fragments within images and across successive images.

To separate infeasible solutions by inequalities (3) for a given t , we label maximal subgraphs of G_t connected by edges labeled 0. Then, for every pair $v, w \in V_t$ of nodes with different labels, we use BFS to search for (i) a shortest vw -path P in $(V_{t+1}, E_{t,t+1} \cup E_{t+1})$ such that $x_P = 0$, and (ii) a vw -cut T in G_t such that $x_T = 1$. We output the inequality defined by P and T . The complexity is $O(|V_t|^2|E_t^+|)$.

To separate infeasible solutions by inequalities (6) for a given t , we start a BFS from every $v \in V_{t+1}$ and by going along edges labeled 0 we either discover some vertex $w \in V_t$ (in this case there is no violation) or the cut $T \in V_t v$ -cuts(G_t^+), which separates v from V_t . In case we found a violation we output the inequality defined by the cut T and vertex v . The complexity is $O(|V_{t+1}| + |E_t^+|)$. The separation of infeasible solutions by inequalities (7) is analogous, in the opposite order of time indices (images).

4.2 Branch-and-Cut

In order to find feasible solutions of the moral lineage tracing problem (Def. 4), with certified bounds, C++ implementations of the separation procedures defined in the previous section are used. Whenever an integer feasible solution is found they are called from the branch-and-cut algorithm of the ILP solver Gurobi [21]. In order to tighten intermediate LP relaxations, we resort to the cuts implemented in Gurobi.

In all experiments we conduct, less than 1% of the total run-time is spent on the separation of infeasible solutions by inequalities (1)–(3), (6) and (7) together. Objective values, bounds and numbers of added inequalities are shown w.r.t. run-time, for three instances of the problem, in Fig. 3.

5 Application to Microscopy Data

In order to examine the effectiveness of the moral lineage tracing problem (MLT) and our proposed branch-and-

cut algorithm as introduced above, we define three instances of the MLT w.r.t. two biomedical data sets, N2DL-HeLa and Flying-Epithelium.

5.1 N2DL-HeLa Data

The N2DL-HeLa data consists of three sequences of images showing HeLa cells that move and divide as bright objects in front of a dark background (Fig. 5a). Two sequences are publicly available and one sequence is undisclosed for an annual competition [33]. Here, we use the two public sequences, one for learning a cost function, the other for experiments (HeLa-test). To additionally obtain a shorter sequence of smaller images, we crop a sub-problem from this test set (HeLa-small). For both test sequences, a priced hypothesis graph is constructed as shown in Fig. 4(a) and described in detail below. The hypothesis graph for the full test set consists of 10882 nodes and 19807 edges. The hypothesis graph for the small, cropped test set consists of 512 nodes and 812 edges.

Optimization. The convergence of the branch-and-cut algorithm for the instances of the MLT w.r.t. the full and the small HeLa data set is shown in the first two rows of Fig. 3. It can be seen from this figure that the small problem is solved to optimality. It can also be seen that the full problem is solved with a certified optimality gap, determined by the lower bound. Finally, it can be seen that most separating cuts are morality constraints.

Results. The lineage forest defined by the solution of the small problem is depicted in Fig. 7. This lineage forest is in exact accordance with the ground truth. The lineage forest defined by the feasible solution of the full problem is depicted in Fig. 8(a). Corresponding decompositions of images are depicted in Fig. 5. Compared to the ground truth provided in [33] in terms of the metrics defined there, the feasible solutions achieve a segmentation accuracy SEG of 0.781 and a tracing accuracy TRA of 0.975. SEG is the average intersec-

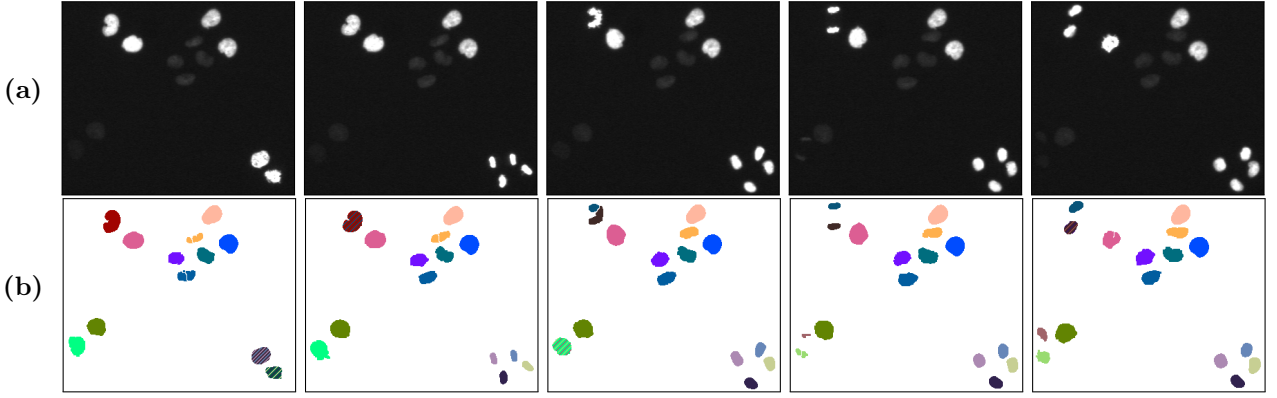


Figure 5 Depicted above are (a) images crops of the full HeLa-test data set, and (b) decompositions of the images defined by a feasible solution of the moral lineage tracing problem (Def. 4). Diagonally striped cells indicate cell division.

tion over union of matched cells. TRA is a weighted edit distance between two lineage forests with weights chosen to reflect the time it takes a human curator to carry out the edit manually.

Technical details. For preprocessing, we train a random forest \mathcal{R}_{FB} to classify pixels as ‘foreground’ or ‘background’. Let $P_F := \mathcal{R}_{FB}(I_{\text{test}}^{\text{HeLa}})$ denote the foreground probability maps for the test image sequence. A binary foreground label image L_F is defined by thresholding P_F at 0.5. We then apply a watershed search on the distance transform of L_F to separate connected components into a set of cell fragments S^{HeLa} . Note that S^{HeLa} denotes the set of all cell fragments for the entire image sequence $I_{\text{test}}^{\text{HeLa}}$. We write S_t^{HeLa} to denote only those cell fragments that correspond to time-point t . A second random forest classifier \mathcal{R}_C is trained to detect cell boundaries, *i.e.* the interface between cells and background, as well as the interface between two touching cells. We denote by $P_C := \mathcal{R}_C(I_{\text{test}}^{\text{HeLa}})$ the probability maps for cell boundaries. Later we will use P_C to determine cut-costs of edges $\bigcup_t E_{t,t+1}$, *i.e.*, edges connecting nodes in adjacent time-points.

The hypothesis graph $G_{\text{HeLa}} := (V, E, \tau)$ for $I_{\text{test}}^{\text{HeLa}}$ is constructed as follows: For each time-point t and each cell fragment $s \in S_t^{\text{HeLa}}$, we introduce a node v_s to V_t with $\tau(v) = t$. For each time-point t and every pair of cell fragments $\{s_i, s_j\} \in (S_t^{\text{HeLa}} \times S_t^{\text{HeLa}})$, we introduce an edge $e_{i,j} \in E_t$ iff the Euclidean distance between the center of mass (COM) of these two cell fragments is less than or equal to a constant d_1 . For each time-point t and every pair of cell fragments $\{s_i, s_j\} \in (S_t^{\text{HeLa}} \times S_{t+1}^{\text{HeLa}})$, we introduce an edge $e_{i,j}$ to $E_{t,t+1}$ iff the distance between their COMs is less than a constant d_2 .

The cost function is defined as follows. All appearance (disappearance) costs are set to the same constant c^+ (c^-). For each t and every $e = \{v_i, v_j\} \in E_{t,t+1}$, we introduce the cut-cost $c_e = \|\text{com}(s_i) - \text{com}(s_j)\|/d_2$ where com denotes the center of mass. For each t and every $e = \{v_i, v_j\} \in E_t$, we introduce the cut-cost $c_e = \max\{\|\text{com}(s_i) - \text{com}(s_j)\|/d_1, b(P_C, \text{com}(s_i), \text{com}(s_j))\}$. Here $b(P_C, \text{com}(s_i), \text{com}(s_j))$ denotes the maximum cut probability in P_C found along the Bresenham line between $\text{com}(s_i)$ and $\text{com}(s_j)$.

5.2 Flywing-Epithelium Data

Images sequences in this data set show a developing fly wing epithelium (Fig. 6a). Here, every pixel is part of a cell and no pixels show background. The data set is divided into a training and test set, denoted by $I_{\text{train}}^{\text{fly}}$ and $I_{\text{test}}^{\text{fly}}$, respectively. We collected ground truth for this data set by manually merging watershed superpixels. The construction of a priced hypothesis graph from the raw test sequence is sketched in Fig. 4b and described in more detail below. It consists of 5026 nodes and 19011 edges.

Optimization. The convergence of the branch-and-cut algorithm for the instance of the MLT for this data set is shown in the third row of Fig. 3. It be seen from this figure that the problem is solved with a certified optimality gap, determined by the lower bound.

Results. The lineage forest defined by the feasible solution of the problem is depicted in Fig. 8(b). Corresponding decompositions of images are depicted in Fig. 6(c). Decompositions and the lineage forests are compared in Tab. 1 to the ground truth in terms of the metrics SEG and TRA. It can be seen from this table that these results are comparable to those found by a state-of-the-art tracking system [2].

Technical details. For preprocessing, we train a random forest classifier \mathcal{R}_I for detecting cell membranes. We denote by $P_I = \mathcal{R}_I(I_{\text{test}}^{\text{fly}})$ the probability map obtained from this classifier. We decompose images into cell fragments S^{fly} by first applying a watershed transform on the raw image sequence $I_{\text{test}}^{\text{fly}}$, and then reduce the large number of watershed segments by progressively merging adjacent superpixels s_i, s_j iff the average intensity value of all pixels in $I_{\text{test}}^{\text{fly}}$ and P_I lying on their interface is below adequately chosen constants t_I and t_P , respectively. We choose those parameters such that under-decompositions is avoided. This leads to 3.09 ± 1.3 superpixels per cell. As cells move considerably between adjacent time-points, we compute a dense optical flow on $I_{\text{test}}^{\text{fly}}$. Let $f(s)$ denote the flow-vector at the COM of a superpixel $s \in S^{\text{fly}}$.

The hypothesis graph $G_{\text{fly}} = (V, E, \tau)$ is constructed as follows: For each time-point t and every cell fragment $s \in S_t^{\text{fly}}$, we introduce a node v_s to V_t with $\tau(v) = t$.

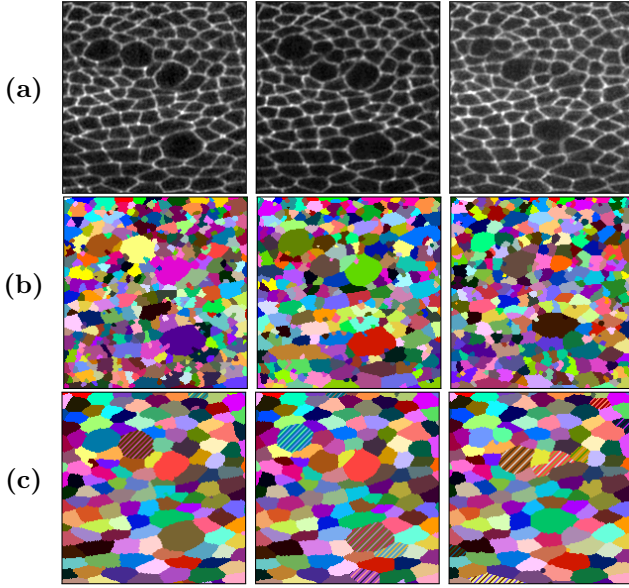


Figure 6 Depicted above are (a) images of the fly wing test set, (b) decompositions of these images into cell fragments, and (c) decompositions of the images defined by a feasible solution of the moral lineage tracing problem (Def. 4). Diagonally striped cells divide in the next image.

For each time-point t and every pair of cell fragments $\{s_i, s_j\} \in S_t \times S_t$, we introduce an edge $e_{i,j}$ to E_t iff s_i is adjacent to s_j . For each time-point t and every pair of cell fragments $\{s_i, s_j\} \in (S_t^{\text{fly}} \times S_{t+1}^{\text{fly}})$, we introduce an edge $e_{i,j}$ to $E_{t,t+1}$ iff $\|(\text{com}(s_i) + f(s_i)) - \text{com}(s_j)\| \leq d$.

The cost function is defined as follows. All appearance (disappearance) costs are set to the same constant c^+ (c^-). For each t and every $e = \{v_i, v_j\} \in E_t$, we introduce the cut-cost $c_e = \frac{\sum_{p \in P(s_i, s_j)} L_B(p)}{|P(s_i, s_j)|}$. For each t and every $e = \{v_i, v_j\} \in E_{t,t+1}$, we introduce the cut-cost $c_e = \frac{1}{1 + \exp(-c_0 - c_1 * m(e))}$, where we values c_0 and c_1 are estimated by logistic regression from the training data. Here, $m(e)$ for $e \in E_t$ denotes the maximum value in $P_{I,t}$ along the geodesic path between the COM of cell fragments $\{s_i, s_j\}$.

6 Conclusion

Building on recent work in image decomposition and people tracking, we have proposed a rigorous mathematical abstraction of lineage tracing, a central problem in biological image analysis. The optimization problem we propose, a hybrid of the well-known minimum cost multicut problem and the minimum cost k disjoint arborescence problem, is a joint formulation of image decomposition and lineage forest reconstruction. Its feasible solutions define, for every image in a sequence of images, a decomposition into cells and, across images, a lineage forest of cells. Unlike previous formulations, it does not constrain the set of decompositions.

We have studied three instances of this problem defined by two biologically relevant microscopy data sets. One instance was solved to global optimality, yielding a solution in exact accordance with decompositions and

Method	SEG	TRA
PA (on GT seg.)	0.9327	0.9898
PA (auto)	0.7980	0.9206
MLT (our)	0.9722	0.9813

Table 1 Quantified above is the distance from ground truth of decompositions (SEG) and traced lineage forests (TRA) as they were obtained algorithmically by moral lineage tracing (MLT) and a state-of-the-art cell tracing system [2].

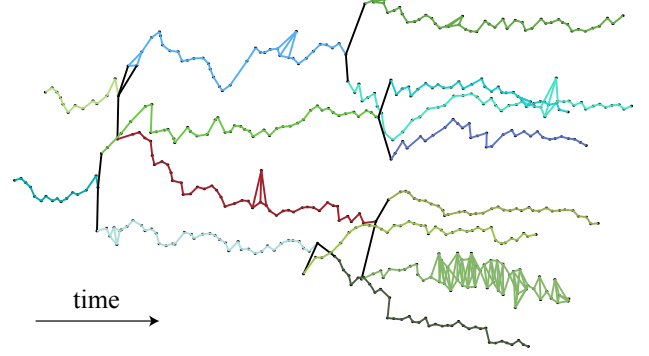


Figure 7 Depicted above is the lineage forest (V, \tilde{C}) reconstructed by solving an instance of the moral lineage tracing problem (Def. 4) defined w.r.t. the image sequence HeLa-small. Edges connecting a fragment of one cell to a fragment of a descendant cell (depicted in black) indicate cell divisions. Edges connecting fragments of the same cell are depicted in a color representing that cell. Note the two progenitor cells in the first image, visible here on the l.h.s..

ground truth lineages. For one larger and harder instances we could obtain feasible solution with a certified optimality gap that compare favorably to decompositions and lineages found by a state-of-the-art lineage tracing system.

Acknowledgements.

We thank the lab of Suzanne Eaton (MPI-CBG) for providing the fly wing data. This work was supported by the German Federal Ministry of Research and Education (BMBF) under the funding code 031A099.

References

- [1] R. Adams and L. Bischof. Seeded region growing. *TPAMI*, 16(6):641–647, June 1994.
- [2] B. Aigouy, R. Farhadifar, D. B. Staple, A. Sagner, J.-C. Röper, F. Jülicher, and S. Eaton. Cell flow reorients the axis of planar polarity in the wing epithelium of *Drosophila*. *Cell*, 142(5):773–786, Sept. 2010.
- [3] A. Alush and J. Goldberger. Ensemble segmentation using efficient integer linear programming. *TPAMI*, 34(10):1966–1977, 2012.
- [4] F. Amat and P. J. Keller. Towards comprehensive cell lineage reconstructions in complex organisms using light-sheet microscopy. *Dev. Growth Differ.*, 55(4):563–578, May 2013.

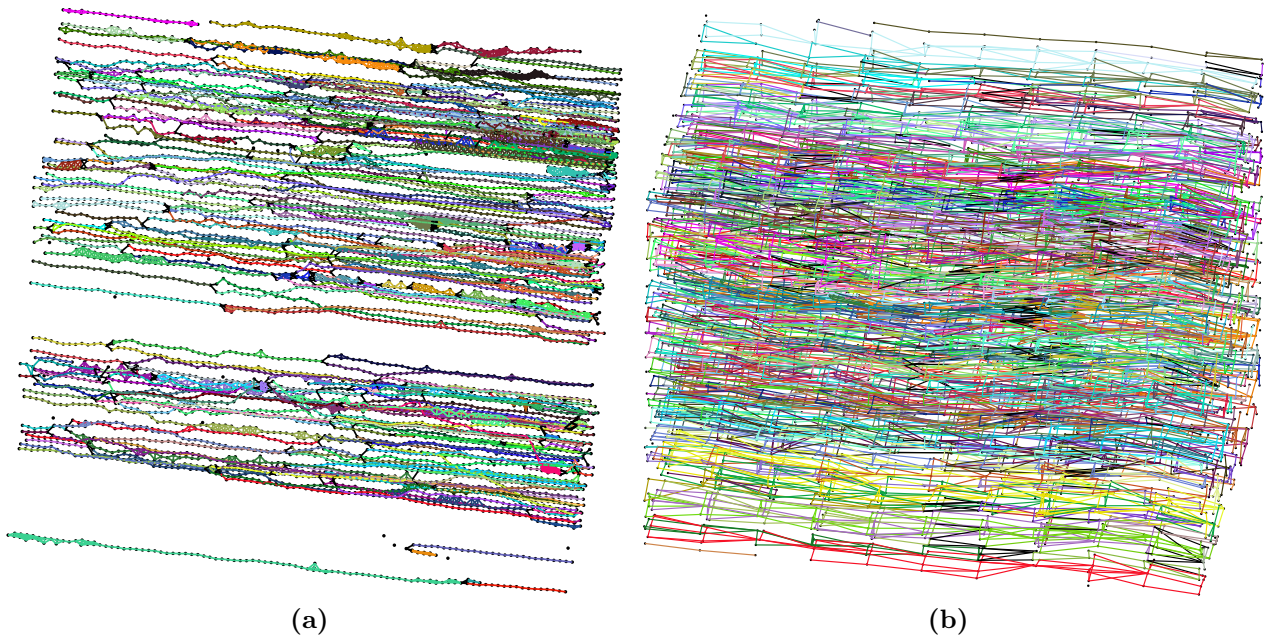


Figure 8 3D rendered lineage forests for (a) the full HeLa-test data set, and (b) the complete fly wing data set, as obtained by solving the moral lineage tracing problem. For better visibility, only the traced moral lineages are shown while all cut edges are hidden. Time progresses from left to right.

- [5] F. Amat, W. Lemon, D. P. Mossing, K. McDole, Y. Wan, K. Branson, E. W. Myers, and P. J. Keller. Fast, accurate reconstruction of cell lineages from large-scale fluorescence microscopy data. *Nature Methods*, 11(9):951–958, Sept. 2014.
- [6] B. Andres. Lifting of multicuts. *CoRR*, abs/1503.03791, 2015.
- [7] B. Andres, J. H. Kappes, T. Beier, U. Köthe, and F. A. Hamprecht. Probabilistic image segmentation with closedness constraints. In *ICCV*, 2011.
- [8] B. Andres, T. Kröger, K. L. Briggman, W. Denk, N. Korogod, G. Knott, U. Köthe, and F. A. Hamprecht. Globally optimal closed-surface segmentation for connectomics. In *ECCV*, 2012.
- [9] B. Andres, J. Yarkony, B. S. Manjunath, S. Kirchoff, E. Türetken, C. C. Fowlkes, and H. Pfister. Segmenting planar superpixel adjacency graphs w.r.t. non-planar superpixel affinity graphs. In *EMMCVPR*, 2013.
- [10] P. Arbeláez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *TPAMI*, 33(5):898–916, 2011.
- [11] S. Bagon and M. Galun. Large scale correlation clustering optimization. *CoRR*, abs/1112.2903, 2011.
- [12] N. Bansal, A. Blum, and S. Chawla. Correlation clustering. *Machine Learning*, 56(1–3):89–113, 2004.
- [13] T. Beier, F. A. Hamprecht, and J. H. Kappes. Fusion moves for correlation clustering. In *CVPR*, 2015.
- [14] T. Beier, T. Kröger, J. H. Kappes, U. Köthe, and F. A. Hamprecht. Cut, Glue & Cut: A fast, approximate solver for multicut partitioning. In *CVPR*, 2014.
- [15] N. Chenouard, I. Smal, F. de Chaumont, M. Maška, I. F. Sbalzarini, Y. Gong, J. Cardinale, C. Carthel, S. Coraluppi, M. Winter, A. R. Cohen, W. J. Godinez, K. Rohr, Y. Kalaidzidis, L. Liang, J. Duncan, H. Shen, Y. Xu, K. E. G. Magnusson, J. Jaldén, H. M. Blau, P. Paul-Gilloteaux, P. Roudot, C. Kervrann, F. Waharte, J.-Y. Tinevez, S. L. Shorte, J. Willemse, K. Celler, G. P. van Wezel, H.-W. Dan, Y.-S. Tsai, C. Ortiz-de Solorzano, J.-C. Olivo-Marin, and E. Meijering. Objective comparison of particle tracking methods. *Nature Methods*, 11(3):281–289, Mar. 2014.
- [16] S. Chopra and M. Rao. The partition problem. *Mathematical Programming*, 59(1–3):87–115, 1993.
- [17] E. D. Demaine, D. Emanuel, A. Fiat, and N. Immerlica. Correlation clustering in general weighted graphs. *Theoretical Computer Science*, 361(2–3):172–187, 2006.
- [18] M. M. Deza and M. Laurent. *Geometry of Cuts and Metrics*. Springer, 1997.
- [19] J. Edmonds. Some well-solved problems in combinatorial optimization. In B. Roy, editor, *Combinatorial Programming: Methods and Applications*, volume 19 of *NATO Advanced Study Institutes Series*, pages 285–301. Springer, 1975.
- [20] J. Funke, B. Anders, F. A. Hamprecht, A. Cardona, and M. Cook. Efficient automatic 3D-reconstruction of branching neurons from EM data. In *CVPR*, 2012.
- [21] I. Gurobi Optimization. Gurobi optimizer reference manual, 2015.

- [22] F. Jug, T. Pietzsch, D. Kainmüller, J. Funke, M. Kaiser, E. van Nimwegen, C. Rother, and G. Myers. Optimal Joint Segmentation and Tracking of Escherichia Coli in the Mother Machine. In *Bayesian and graphical Models for Biomedical Imaging*, pages 25–36. Springer, Cham, 2014.
- [23] J. H. Kappes, M. Speth, B. Andres, G. Reinelt, and C. Schnörr. Globally optimal image partitioning by multicuts. In *EMMCVPR*, 2011.
- [24] J. H. Kappes, M. Speth, G. Reinelt, and C. Schnörr. Higher-order segmentation via multicuts. *CoRR*, abs/1305.6387, 2013.
- [25] J. H. Kappes, P. Swoboda, B. Savchynskyy, T. Hazan, and C. Schnörr. Probabilistic correlation clustering and image partitioning using perturbed multicuts. In *SSVM*, 2015.
- [26] B. X. Kausler, M. Schiegg, B. Andres, M. Lindner, U. Koethe, H. Lette, J. Wittbrodt, L. Hufnagel, and F. A. Hamprecht. A discrete chain graph model for 3d+t cell tracking with high misdetection robustness. In *ECCV*, 2012.
- [27] P. J. Keller. Imaging morphogenesis: technological advances and biological insights. *Science*, 340(6137):1234168, June 2013.
- [28] P. J. Keller, A. D. Schmidt, A. Santella, K. Khairy, Z. Bao, J. Wittbrodt, and E. H. K. Stelzer. Fast, high-contrast imaging of animal development with scanned light sheet-based structured-illumination microscopy. *Nature Methods*, 7(8):637–642, Aug. 2010.
- [29] P. J. Keller, A. D. Schmidt, J. Wittbrodt, and E. H. K. Stelzer. Reconstruction of zebrafish early embryonic development by scanned light sheet microscopy. *Science*, 322(5904):1065–1069, Nov. 2008.
- [30] M. Keuper, E. Levinkov, N. Bonneel, G. Lavoué, T. Brox, and B. Andres. Efficient decomposition of image and mesh graphs by lifted multicuts. In *ICCV*, 2015.
- [31] K. Khairy and P. J. Keller. Reconstructing embryonic development. *Genesis*, 49(7):488–513, July 2011.
- [32] S. Kim, C. Yoo, S. Nowozin, and P. Kohli. Image segmentation using higher-order correlation clustering. *TPAMI*, 36:1761–1774, 2014.
- [33] M. Maška, V. Ulman, D. Svoboda, P. Matula, P. Matula, C. Ederra, A. Urbiola, T. España, S. Venkatesan, D. M. W. Balak, P. Karas, T. Bolcková, M. Štreitová, C. Carthel, S. Coraluppi, N. Harder, K. Rohr, K. E. G. Magnusson, J. Jaldén, H. M. Blau, O. Dzyubachyk, P. Křížek, G. M. Hagen, D. Pastor-Escuredo, D. Jimenez-Carretero, M. J. Ledesma-Carbayo, A. Muñoz-Barrutia, E. Meijering, M. Kozubek, and C. Ortiz-de Solorzano. A benchmark for comparison of cell tracking algorithms. *Bioinformatics*, 30(11):1609–1617, 2014.
- [34] S. G. Megason and S. E. Fraser. Imaging in systems biology. *Cell*, 130(5):784–795, Sept. 2007.
- [35] D. Padfield, J. Rittscher, and B. Roysam. Coupled minimum-cost flow cell tracking for high-throughput quantitative analysis. *Medical Image Analysis*, 15(4):650–668, Aug. 2011.
- [36] M. Schiegg, P. Hanslovsky, C. Haubold, U. Koethe, L. Hufnagel, and F. A. Hamprecht. Graphical Model for Joint Segmentation and Tracking of Multiple Dividing Cells. *Bioinformatics*, page btu764, Nov. 2014.
- [37] M. Schiegg, P. Hanslovsky, B. X. Kausler, and L. Hufnagel. Conservation Tracking. *ICCV 2013*, 2013.
- [38] A. Schrijver. *Combinatorial optimization. Polyhedra and efficiency*. Springer, 2003.
- [39] J. Shi and J. Malik. Normalized cuts and image segmentation. *TPAMI*, 22(8):888–905, Aug. 2000.
- [40] S. Tang, B. Andres, M. Andriluka, and B. Schiele. Subgraph decomposition for multi-target tracking. In *CVPR*, 2015.
- [41] R. Tomer, K. Khairy, F. Amat, and P. J. Keller. Quantitative high-speed imaging of entire developing embryos with simultaneous multiview light-sheet microscopy. *Nature Methods*, 9(7):755–763, July 2012.
- [42] E. Türetken, C. Becker, P. Glowacki, F. Benmansour, and P. Fua. Detecting irregular curvilinear structures in gray scale and color imagery using multi-directional oriented flux. In *ICCV*, 2013.
- [43] E. Türetken, F. Benmansour, B. Andres, H. Pfister, and P. Fua. Reconstructing loopy curvilinear structures using integer programming. In *CVPR*, 2013.
- [44] E. Türetken, F. Benmansour, and P. Fua. Automated reconstruction of tree structures using path classifiers and mixed integer programming. In *CVPR*, 2012.
- [45] E. Türetken, G. Gonzalez, C. Blum, and P. Fua. Automated reconstruction of dendritic and axonal trees by global optimization with geometric priors. *Neuroinformatics*, 9:279–302, 2011.
- [46] X. Wang, E. Türetken, F. Fleuret, and P. Fua. Tracking interacting objects optimally using integer programming. In *ECCV*, 2014.
- [47] J. Yarkony and C. Fowlkes. Planar ultrametric rounding for image segmentation. In *NIPS*, 2015.
- [48] J. Yarkony, A. Ihler, and C. C. Fowlkes. Fast planar correlation clustering for image segmentation. In *ECCV*, 2012.